

TITLE OF THE INVENTION: METHOD AND SYSTEM FOR DATA PROCESSING AND ANALYSIS

TECHNICAL FIELD

[0001] The present disclosure relates to a method and system for data processing and analysis which utilize artificial intelligence (AI) model.

BACKGROUND

[0002] With the advent of the era of big data, data processing and analysis methods and systems are widely used in finance, banking, e-commerce, and other fields. For enterprises in these fields, data generated is growing exponentially with time. To facilitate better enterprise development, it is necessary to process and analyse the massive data generated during the operation of enterprises based on data processing and analysis methods, so as to provide an effective basis for decision-making by corporate leaders.

[0003] Existing data processing and analysis methods and systems are fraught with issues.

[0004] Typically, data analysts of enterprises manually process and analyse data generated during the operation of these enterprises. As this is time-consuming and highly dependent on individual expertise of the data analysts, data processing efficiency is low and quality of results can vary widely.

[0005] Data between various departments or systems within an enterprise may be independent and lack effective integration and consolidation mechanisms. This leads to prominent data silo problems and difficulty in forming a complete data overview.

[0006] Data format and structure of different data sources may vary, hence requiring standardization for integration and analysis. However, this standardization is often time-consuming, laborious, and prone to errors.

[0007] With an exponential increase of enterprise data volume, traditional data processing methods are no longer able to meet the processing needs of large-scale data. This leads to low efficiency in large-scale data processing and inability to provide timely support for decision-making.

[0008] Big data analysis requires the use of complex technologies such as Hadoop, Spark, etc. These technologies have a high technical threshold for general data analysts and require certain technical resources and training.

[0009] Data analysis often requires cross-departmental collaboration, integrating data and resources from multiple departments. However, the organizational division of labor in many

enterprises may not be suitable for the needs of big data analysis, leading to poor collaboration.

[0010] Accordingly, there is a need for data processing and analysis method and systems which address the above and other issues.

SUMMARY

[0011] In an aspect of the disclosure, a computer-implemented data processing and analysis method is provided and comprises:

- performing iterations of data processing and analysis, wherein each iteration comprises:

- receiving a problem statement sent by a terminal device, wherein the problem statement defines a data retrieval requirement corresponding thereto;

- generating a data query based on the data retrieval requirement, and determining a data statistical analysis function based on the problem statement;

- retrieving data based on the data query;

- generating a data statistical analysis result based on the retrieved data and the data statistical analysis function;

- generating, by a trained data analysis model, an anomaly analysis and a trend analysis based on the data statistical analysis result and a second instruction; and

- providing the data statistical analysis result, the anomaly analysis, and the trend analysis to the terminal device or an other device.

[0012] In an embodiment, the step(s) of generating the data query based on the data retrieval requirement and determining the data statistical analysis function based on the problem statement includes:

- generating, using the trained data analysis model, the data query based on the data retrieval requirement and a first instruction; and

- generating, using the trained data analysis model, the data statistical analysis function based on the problem statement and the first instruction.

[0013] In an embodiment, the method further comprises:

- using first training datasets, performing iterations of training of an initial data analysis model to obtain an intermediate data analysis model, wherein each first training dataset includes first training model inputs and first training model outputs, wherein the first training model inputs include a training problem statement, and a first training instruction, wherein first training model outputs include a training data query, and a training data statistical analysis function; and

using second training datasets, performing iterations of training of the intermediate data analysis model to produce the trained data analysis model, wherein each second training dataset includes second training model inputs and second training model outputs, wherein the second training model inputs include a training data statistical analysis result, and a second training instruction, wherein the second training model outputs include a training anomaly analysis and a training trend analysis.

[0014] In an embodiment, the step of performing the iterations of training of the initial data analysis model to obtain the intermediate data analysis model includes:

after performing each iteration of training:

determining a loss function of the initial data analysis model;

determining the loss function of the initial data analysis model is non-convergent and adjusting model parameters of the initial data analysis model;

wherein the method further comprises:

stop performing the iterations of training of the initial data analysis model if a cumulative training duration of the initial data analysis model exceeds a first predetermined cumulative training duration threshold, or a cumulative training iteration count of the initial data analysis model exceeds a first predetermined cumulative training iteration count threshold, or a validation of the initial data analysis model exceeds a first predetermined validation requirement; and

wherein performing the iterations of training of the intermediate data analysis model to produce the trained data analysis model includes:

after performing each iteration of training:

determining a loss function of the intermediate data analysis model;

determining the loss function of the intermediate data analysis model is non-convergent and adjusting model parameters of the intermediate data analysis model;

wherein the method further comprises:

stop performing the iterations of training of the intermediate data analysis model if a cumulative training duration of the intermediate data analysis model exceeds a second predetermined cumulative training duration threshold, or a cumulative training iteration count of the intermediate data analysis model exceeds a second predetermined cumulative training iteration count threshold, or a validation of the intermediate data analysis model exceeds a second predetermined validation requirement.

[0015] In an embodiment, the method further comprises:

retraining the trained data analysis model, when a cumulative count of the iterations of data processing and analysis exceeds a predetermined cumulative count of iterations of data processing and analysis, to obtain a retrained data analysis model including:

using first precedent datasets obtained from performing the iterations of data processing and analysis, performing iterations of retraining of the trained data analysis model to obtain an intermediately-retrained data analysis model, wherein each first precedent dataset includes first precedent model inputs and first precedent model outputs, wherein the first precedent model inputs include the problem statement, and the first instruction, wherein first precedent model outputs include the data query, and the data statistical analysis function;

using second precedent datasets obtained from performing iterations of data processing and analysis, performing iterations of retraining on the trained data analysis model to obtain an intermediately-retrained data analysis model wherein each second precedent dataset includes second precedent model inputs and second precedent model outputs, wherein the second precedent model inputs include the data statistical analysis result, and the second instruction, wherein the second precedent model outputs include the anomaly analysis and the trend analysis; and

using the retrained data analysis model, updating the trained data analysis model.

[0016] In an embodiment, the method further comprises:

receiving a correction request having at least one of the group consisting of a corrected data query, a corrected data statistical analysis function, a correct data statistical analysis result, a corrected anomaly analysis, and a corrected trend analysis;

storing the correction request; and

retraining the trained data analysis model based on the correction request, wherein at least one of the first precedent datasets or at least one of the second precedent datasets is based on the correction request.

[0017] In an embodiment, the step of generating the data query based on the data retrieval requirement includes:

generating a Structured Query Language (SQL) statement; and

retrieving the data from the database based on the generated SQL statements.

[0018] In an embodiment, the step of generating the data statistical analysis result based on the retrieved data and the data statistical analysis function includes:

storing a plurality of sample problem statements and a plurality of sample data statistical analysis functions corresponding to the sample problem statements;

successfully matching the problem statement to one of the sample problem statements;

determining a corresponding sample data statistical analysis function of the matched sample problem statement as the data statistical analysis function corresponding to the problem statement.

[0019] In an embodiment, the step of generating the data query based on the data retrieval requirement includes:

determining a domain knowledge of the problem statement; and

generating, using the data analysis model, the data query based on the problem statement based on the domain knowledge and the problem statement.

[0020] In an embodiment, the step of determining the domain knowledge of the problem statement includes:

determining similarity scores of sample problem statements relative to the problem statement, determining a corresponding one of the sample problem statements as having a similarity score higher than a predetermined score threshold, determining a domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement.

[0021] In an embodiment, the step of determining the domain knowledge of the problem statement includes:

determining keywords of the problem statement, determining matching counts of the sample problem statements relative to the problem statement, determining a corresponding one of the sample problem statements as having a highest matching count, determining a domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement.

[0022] In an embodiment, the step of generating the data statistical analysis result based on the retrieved data and the data statistical analysis function includes:

substituting the retrieved data into the data statistical analysis function; and

performing the data statistical analysis function to generate the data statistical analysis result.

[0023] In an aspect of the disclosure, a non-transitory computer-readable medium is provided and comprises computer-executable instructions configured to direct at least one computer processor to perform the method according to any one of the preceding aspects and/or embodiments.

[0024] In an aspect of the disclosure, a data processing and analysis device is provided and comprises:

at least one memory for storing computer-executable instructions; and

at least one processor communicably coupled to the at least one memory, and configured to:

execute the computer-executable instructions to perform the method according to any one of the preceding aspects and/or embodiments.

BRIEF DESCRIPTION OF DRAWINGS

Embodiments of the invention will be described in detail with reference to the accompanying drawings, in which:

[0025] Figure 1 shows a data processing and analysis method according to an embodiment;

[0026] Figures 2A and 2B show data processing and analysis methods according to some embodiments;

[0027] Figure 3 shows a data processing and analysis method according to an embodiment; and

[0028] Figure 4 shows a schematic representation of a system for data processing and analysis.

DETAILED DESCRIPTION

[0029] In the following description, numerous specific details are set forth in order to provide a thorough understanding of various illustrative embodiments of the invention. It will be understood, however, to one skilled in the art, that embodiments of the invention may be practiced without some or all of these specific details. It is understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the invention. In the drawings, like reference labels or numerals refer to same or similar functionalities or features throughout the several views. In the drawings, directional arrows shown between features illustrate data or information transfer therebetween in accordance with description of some embodiments but are not limited as such. In other words, data or information transfer in reverse to the directional arrows and/or not shown by directional arrow among features may be envisaged and are not shown to avoid obscuring description of the embodiments.

[0030] Embodiments described in the context of one of the apparatuses or methods are analogously valid for the other apparatuses or methods. Similarly, embodiments described in the context of an apparatus are analogously valid for a method, and vice versa.

[0031] Features that are described in the context of an embodiment or example may correspondingly be applicable to the same or similar features in the other embodiments or examples. Features that are described in the context of an embodiment or example may correspondingly be applicable to the other embodiments or examples, even if not explicitly described in these other embodiments or examples. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment or example may correspondingly be applicable to the same or similar feature in the other embodiments or examples.

[0032] It should be understood that the articles "a", "an" and "the" as used with regard to a feature or element include a reference to one or more of the features or elements. The term "and/or" includes any and all combinations of one or more of the associated feature or element. The terms "comprising", "including", "having", "involving" and any of their related terms, as used in description and claims, are intended to be open-ended and mean that there may be additional features or elements other than the listed ones. Identifiers such as "first", "second", "third", and so on, are used merely as labels, and are not intended to impose numerical requirements on their objects, nor construed in a manner imposing any relative position or time sequence between limitations.

[0033] Figure 1 shows a flowchart relating to a data processing and analysis method according to an embodiment.

[0034] In the flowchart of Figure 1, block 100 includes receiving a problem statement including a data retrieval requirement corresponding to the problem statement.

[0035] The problem statement defines one or more issues or topics to be analyzed, and may be defined by a human user or computer, in a structured or non-structured format. The data retrieval requirement defines target data required for analyzing the issue(s). Non-limiting examples of data retrieval requirement may include target department, target time period, and/or target metrics. In other words, based on the problem statement, target departmental data within the target enterprise, target metrics, and/or target time period which are required to analyse the issue can be determined.

[0036] For example, for a target enterprise which is an e-commerce company, a problem statement may be Issue A—analyze sales performance of Sales Department A in July 2024. Based on Issue A, a data retrieval requirement which identifies the target data needed for analyzing the issue can be determined, i.e., the target department is Sales Department A, the target metrics are monthly total sales, daily average sales, highest single-day sales, lowest single-day sales, etc., and the target time period is from July 1, 2024, to July 31, 2024.

[0037] For example, for a target enterprise which is a bank or financial institution, a problem statement may be Issue B—analyze credit card processing situation of Department B in May 2024. Based on Issue B, a data retrieval requirement which identifies the target data needed for analyzing the issue can be determined, i.e., the target department is Department B, the target metrics are monthly total number of credit cards processed, daily average number of credit cards processed, highest single-day number of credit cards processed, lowest single-day number of credit cards processed, etc., and the target time period is from May 1, 2024, to May 31, 2024.

[0038] Block 100 may be performed by using a computer processor at a data analysis device, e.g., Figure 4, upon instruction by a data processing application, to receive an analysis request which includes a problem statement and is sent directly or indirectly by a terminal device or

an other device. When user(s) at the terminal device need an analysis of the data generated during operation of the target enterprise, they may compile issues to be analyzed based on actual needs and send analysis request(s) from the terminal device to the data analysis device, where the data processing application may have been installed through the terminal device.

[0039] In the flowchart of Figure 1, block 200 includes generating a data query based on the data retrieval requirement and determining a data statistical analysis function based on the problem statement; and retrieving data based on the data query.

[0040] The data query includes instructions or commands for performing tasks on database(s). Particularly, the data query includes instructions at least for selectively retrieving target data according to the data retrieval requirement of the problem statement. Based on the data query, target data from one or more database(s) is retrieved.

[0041] For example, Structured Query Language (SQL) statement may be generated for searching the data to be analyzed, and the generated SQL statement may be performed for querying the database(s) to obtain the data to be analyzed. Alternative to SQL, other data query approaches or languages, e.g. Microsoft Access, PostgreSQL, Oracle Database, Pipelined Relational Query Language (PRQL), Graph Query Language (GQL), may be used instead.

[0042] Data statistical analysis function(s) corresponding to issue(s) to be analyzed may be determined prior to, after, or at the time of data retrieval. Non-limiting examples of data statistical analysis function include operations for summation, averaging, finding the maximum value, finding the minimum value based on certain target data.

[0043] For example, multiple sample problem statements and their corresponding data statistical analysis functions may be stored in a database beforehand, and the problem statement(s) to be analyzed are matched with the multiple stored sample problem statements. If the problem statement(s) to be analyzed are successfully matched with one or more stored sample problem statement(s), the corresponding stored data statistical analysis function(s) may be determined as the data statistical analysis function(s) corresponding to the problem statement(s) to be analyzed.

[0044] For example, an AI model may generate the data query and/or data statistical analysis function(s) based on a given problem statement.

[0045] Block 200 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to generate the data query, retrieve target data from one or more databases, and determine the data statistical analysis function from the given problem statement.

[0046] In the flowchart of Figure 1, block 300 includes generating a data statistical analysis result based on the retrieved data and the data statistical analysis function.

[0047] For example, the retrieved data may be substituted into the data statistical analysis function(s), and the data statistical analysis function(s) are executed or performed to generate or output a data statistical analysis result corresponding to the given problem statement.

[0048] For example, an AI model may generate data statistical analysis result(s) based on the data statistical analysis.

[0049] Block 300 may be performed by using the computer processor of the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to generate the data statistical analysis.

[0050] In the flowchart of Figure 1, block 400 includes generating, by a trained data analysis model, an anomaly analysis and a trend analysis based on the data statistical analysis result and a preset instruction, e.g., second preset instruction. The preset instruction, e.g., preset prompt, is configured to guide the data analysis model to carry out a specific task, e.g., generate required output(s) based on prescribed input(s).

[0051] The trained data analysis model, may be obtained through iterative training of a predetermined large model using training datasets. The predetermined large model may be any Large Language Model (LLM) that may be known in the field of data processing.

[0052] Each training dataset includes training model inputs and training model outputs. The training model inputs include a training data statistical analysis result and a (second) preset instruction, while the training model outputs include a training anomaly analysis result and a training trend analysis result corresponding to the training data statistical analysis result. Here, the reference label “training” generally refers to parameters used for training the data analysis model in accordance with block 400.

[0053] Since a plurality of training model inputs are used as inputs to the predetermined large model, and the training model outputs corresponding to each training model input are used as outputs of the predetermined large model, the predetermined large model is iteratively trained to obtain the trained data analysis model. After training, the trained data analysis model is configured to receive inputs, including a data statistical analysis result and a (second) preset instruction, and generate outputs, including a anomaly analysis result and a trend analysis result corresponding to the inputted data statistical analysis result.

[0054] The anomaly analysis result corresponding to the given problem statement may include relevant information on the retrieved data with anomalies in the data to be analyzed, such as date of the anomalous data, anomaly situation of the anomalous data, cause analysis of the anomalous data, etc.

[0055] For example, if the problem statement is sales performance of Sales Department A in July 2024, and the data to be analyzed is the daily sales of each employee in the Sales Department A in July 2024, non-limiting examples of anomaly analysis results corresponding to the issue to be analyzed may include : (i) Employee X's sales on July 23, 2024 are

significantly higher than on other dates, and the corresponding reason for this anomaly; (ii) Sales Department A's sales on July 3, 2024, are significantly lower than on other dates, and the corresponding reason for this anomaly, etc.

[0056] The trend analysis results corresponding to the given problem statement may include trends and patterns of change corresponding to the retrieved data. Non-limiting examples of trend analysis include Time Series Analysis for observing trends in data over time, such as linear growth, exponential growth, or seasonal fluctuations (this helps identify long-term trends and seasonal variations in the data); Correlation Analysis for exploring the correlations between different variables to determine which factors may jointly influence the data's trend (for example, in sales data analysis, a correlation might be found between sales revenue, advertising expenditures, seasonal changes, or consumer preferences; Cluster Analysis for grouping data into similar categories or clusters to identify patterns and outliers within the data (this helps uncover hidden structures in the data, such as specific customer segments or market niches); Regression Analysis for using statistical models to predict how one variable, e.g., sales revenue, changes with variations in other variables, e.g., price, advertising expenditures, (this helps quantify the impact of different factors on the target variable and predict future trends); Visual Analysis for utilizing visualization tools such as charts, graphs, and dashboards to display trends and patterns in the data (this aids in intuitively understanding the data and discovering subtle changes that may be difficult to detect through purely numerical analysis).

[0057] Block 400 may be performed by using the computer processor of the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to input a data statistical analysis result and a second preset instruction into the trained data analysis model, and obtain model outputs therefrom, including an anomaly analysis and a trend analysis corresponding to the inputs.

[0058] In the flowchart of Figure 1, block 500 includes providing the data statistical analysis result, the anomaly analysis, and the trend analysis to the terminal device or an other device.

[0059] This allows the terminal device or the other device to output and display the data statistical analysis result, anomaly analysis result, and trend analysis result corresponding to the given problem statement. Subsequently, users at the terminal device or the other device may make decisions based on the provided or displayed information.

[0060] Block 500 may include providing and/or displaying, by a display device communicably coupled to the terminal device or the other device, the data statistical analysis result, the anomaly analysis, and the trend analysis, the analysis request, the problem statement, the retrieved data, the data query, and/or the data statistical analysis function.

[0061] With the displayed information, the user may analyze whether there are issues with the data query, data statistical analysis function, anomaly analysis result, and trend analysis

result. If the user determines that there is any issue, the user may accordingly generate and/or send a correction request including a corrected version, e.g., at least one of the group consisting of a corrected data query, a corrected data statistical analysis function, a corrected data statistical analysis result, a corrected anomaly analysis, and a corrected trend analysis. Depending on the nature of correction, blocks 200, 300 and/or 400 may be performed based on the corrected version.

[0062] For example, if the problem statement is corrected in the correction request, the correction request may include a corrected data retrieval requirement. Then, according to block 200, a resulting corrected data query is generated based on the resulting corrected data retrieval requirement of the corrected problem statement; resulting corrected data corresponding to the corrected data query is retrieved; and a resulting corrected data statistical analysis function based on the corrected problem statement is determined based on the correction request. Subsequently, according to block 300, a resulting corrected data statistical analysis result is generated by applying the retrieved corrected data to the corrected data statistical analysis function. Subsequently, according to block 400, using the trained data analysis model, a model output based on the corrected data statistical analysis result is generated which includes a resulting corrected anomaly analysis and a resulting corrected trend analysis. The above-mentioned corrected version, e.g., one or more of the following: a corrected problem statement, a corrected data query, a corrected data retrieval requirement, a corrected data, a corrected data statistical analysis function, a corrected data statistical analysis result, a corrected anomaly analysis, a corrected trend analysis, may be stored in a database for subsequent use in retraining the trained data analysis model to improve accuracy of the output results of the data analysis model. Subsequently, according to block 500, the corrected version may be provided to the terminal device or the other device for display.

[0063] For example, if the anomaly analysis and/or trend analysis are corrected in the correction request, blocks 200, 300 and/or 400 may not be performed based on the correction request. Instead, according to block 500, the corrected version of the anomaly analysis and/or trend analysis may be provided to the terminal device or the other device for display.

[0064] Block 500 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to provide the data statistical analysis result and the model outputs to the terminal device and, optionally, control the terminal device or a display device communicably coupled thereto to display the provided information. Block 500 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon receiving a correction request and instruction by the data processing application, perform some or all of the data processing and analysis step of blocks 200, 300 and 400 based on the corrected version included in the correction request.

[0065] It is to be appreciated that variations and/or modifications to the embodiment of Figure 1 may be envisaged in other embodiments.

[0066] In an embodiment, block 200 of Figure 1 may be performed using an AI model. Particularly, as shown in Figure 2A, in block 200, the step(s) of generating the data query based on the data retrieval requirement and determining the data statistical analysis function based on the problem statement may include generating, using an AI model, the data query based on the data retrieval requirement and a first preset instruction; and generating, using an AI model, the data statistical analysis function based on the problem statement and the first preset instruction. Such AI model may be the trained data analysis model used in block 400 or another AI model(s).

[0067] Block 200 of Figure 2A may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to input the problem statement, the first preset instruction, and data retrieval requirement into the trained data analysis model and/or other model(s), and obtain model outputs therefrom, including the data query and the data statistical analysis function.

[0068] In an embodiment, block 200 of Figure 1 may include further steps. Particularly, as shown in Figure 2B, in block 200, the step of generating the data query based on the data retrieval requirement includes determining a domain knowledge of the problem statement (block 210); and generating, using the trained data analysis model, the data query and data statistical analysis function based on the domain knowledge and the problem statement (block 220).

[0069] In block 210, a plurality of sample problem statements, sample domain knowledge corresponding to each sample problem statement, and a plurality of keywords corresponding to the sample problem statements may be stored in a database. The domain knowledge corresponding to the problem statement may be determined using one of the two following methods.

[0070] In a first method, block 210 includes determining similarity scores of a plurality of sample problem statements relative to the problem statement. Block 210 further includes determining a corresponding one of the sample problem statements as having a similarity score higher than a predetermined score threshold. Block 210 further includes determining a domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement. Particularly, the highest similarity score together with its corresponding sample problem statement and corresponding domain knowledge is identified. The corresponding domain knowledge is then determined as the domain knowledge of the problem statement.

[0071] In a second method, block 210 includes determining keywords of the problem statement, determining matching counts, e.g. number or count of matches, of the sample

problem statements relative to the problem statement, determining a corresponding one of the sample problem statements as having a highest matching count, e.g., highest degree of matching, determining a domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement.

[0072] For example, if Problem Statement A has five keywords, wherein three of them are the same as three keywords of the problem statement to be analyzed, while Problem Statement B has five keywords, wherein four of them are the same as four keywords of the problem statement at hand. As Problem Statement B has a higher matching degree with the problem statement to be analyzed, the domain knowledge corresponding to the successfully matched analyzed problem, i.e., Problem B, is determined as the domain knowledge corresponding to the problem statement to be analyzed.

[0073] In block 220, the problem statement, the identified domain knowledge, and the first preset instruction are inputted into the trained data analysis model, so that the trained data analysis model can output the data query, e.g., SQL statement, and data statistical analysis function corresponding to the problem statement to be analyzed.

[0074] In block 230, using the identified data query, a database can be queried to provide data to be analyzed corresponding to the problem statement.

[0075] Block 200 of Figure 2B may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to determine a domain knowledge of the problem statement; generate, using the trained data analysis model, the data query and data statistical analysis function based on the domain knowledge and the problem statement; and query a database based on the data query.

[0076] In an embodiment, the data analysis model may be pre-trained. Particularly, as shown in Figure 3, block 90, which may precede block 100 to block 500, includes training the data analysis model.

[0077] Training may start with a predetermined large model as an initial data analysis model, then fine-tune this large model using first training datasets to obtain an intermediate data analysis model, and finally train this intermediate or fine-tuned data analysis model with second training datasets. In other words, the trained data analysis model is not directly obtained by training the initial model in the traditional manner. Instead, it is based on the fine-tuned or intermediate model obtained from training the initial model, and then further training is performed on this intermediate model to obtain the trained data analysis model. Using this data analysis model to analyze data not only provides the final output of anomaly analysis results and trend analysis results for the data to be analyzed, but also yields intermediate features, e.g., SQL statements and data statistical analysis functions corresponding to the problem statements, which are to be used for statistical analysis of the retrieved data. Each first training dataset includes first training model inputs which in turn include a training problem

statement, a first training preset instruction, and a training domain-specific knowledge corresponding to the problem statement; and first training model outputs which in turn include a training SQL statement, a training data statistical analysis function corresponding to the training problem statement. Each second training dataset includes second training model inputs which in turn includes a training data statistical analysis result and a second training preset instruction; and second model outputs which in turn include a training anomaly analysis result and a training trend analysis result corresponding to the training data statistical analysis result. The trained data analysis model is obtained by iteratively training an initial data analysis model with the first training datasets to obtain a fine-tuned or intermediate data analysis model, and then iteratively training this fine-tuned or intermediate data analysis model with the second training datasets to obtain the trained data analysis model. Accordingly, the trained data analysis model has two functions. Function 1: When a problem statement is inputted into the trained data analysis model, the model can accurately output the corresponding SQL statement and data statistical analysis function for the problem statement. Function 2: When data statistical analysis result is inputted into the trained data analysis model, the model can accurately output the corresponding anomaly analysis result and trend analysis results for the data statistical analysis result. Due to the complex structure and massive parameters of large models, their size is significant. Therefore, training a single data analysis model with two functions, by using a large model, requires fewer resources and parameters than training two separate large models based on the first and second training samples, respectively. Specifically, if the first large model is iteratively trained with multiple first training datasets to obtain Model A with Function 1, and the second large model is iteratively trained with multiple second training datasets to obtain Model B with Function 2, the combined size of Model A and Model B would be significantly larger than the above-mentioned single data analysis model. Moreover, resource consumption, e.g., memory, GPU computing power, electricity, etc., of Model A and Model B on electronic devices would also be significantly higher than that of the above-mentioned single data analysis model. In some other methods, simpler and smaller models with fewer parameters may be used for training. However, the accuracy of the output results from these trained small models is much lower than that of trained large models. Therefore, if the first small model is iteratively trained with multiple first training datasets to obtain Model C with Function 1, and the second small model is iteratively trained with multiple second training datasets to obtain Model D with Function 2, the accuracy of the output results from Model C and Model D would be much lower than that of the above-mentioned single data analysis model. In summary, training a single data analysis model with two functions, using a large model, balances multiple aspects such as model size, resource utilization, and accuracy of data analysis, thus providing more accurate results while consuming fewer resources.

[0078] Block 90 may include iteratively training a predetermined data analysis model, e.g., large model, based on the first training datasets and the second training datasets to obtain a trained data analysis model. Block 90 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to train an initial data analysis model to obtain a trained data analysis model.

[0079] Particularly, using the first training datasets, block 90 includes using first training datasets, performing iterations of training of an initial data analysis model to obtain an intermediate data analysis model (block 92); and using second training datasets, performing iterations of training of the intermediate data analysis model to obtain a trained data analysis model (block 93).

[0080] Block 90 may further include block 91 which includes obtaining a plurality of first training datasets and a plurality of second training datasets. Each first training dataset may include first training model inputs and first training model outputs. The first training model inputs include a training problem statement, a first training preset instruction, and a training domain knowledge corresponding to the training problem statement. The first training model outputs include a training SQL statement and a training data statistical analysis function corresponding to the training problem statement. The first training preset instruction includes a prompt for the model to output training SQL statement and training data statistical analysis function. The training domain knowledge corresponding to the training problem statement refers to background knowledge of the professional domain involved in the training problem statement. Each second training dataset includes second training model inputs and second training model outputs. The second training model inputs include a training data statistical analysis result and a second preset training instruction. The second training model outputs include a training abnormal analysis result and a training trend analysis results corresponding to the training data statistical analysis result. The second training preset instruction includes a prompt for the model to output abnormal analysis result and trend analysis result.

[0081] Block 91 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to obtain the first training datasets and the second training datasets.

[0082] Block 92 includes, using the first training datasets, performing iterations of training of an initial data analysis model to obtain an intermediate data analysis model. In other words, the iterations of training include using the first training model inputs as inputs of the initial data analysis model and using the first training model outputs corresponding to the first training model inputs as outputs of the initial data analysis model to iteratively train the initial data analysis model.

[0083] Block 92 includes, after each iteration of training, determining whether a loss function of the initial data analysis model converges. If the loss function does not converge, the model

parameters of the initial data analysis model are optimized and adjusted according to the loss function, and the next iteration of training is performed based on the optimized and adjusted data analysis model. If the loss function converges, e.g., after multiple iterations of training, the data analysis model obtained after completion of this latest iteration training is determined as the fine-tuned or intermediate data analysis model.

[0084] It should be noted that since multiple first training model inputs, e.g., problem statements, first preset instructions, and training domain knowledge corresponding to the problem statements, are used as inputs of the data analysis model, and the first training model outputs, e.g., training SQL statements and training data statistical analysis functions corresponding to the training problem statements, corresponding to the first training model inputs are used as outputs of the data analysis model to iteratively train the data analysis model, a fine-tuned or intermediate data analysis model is accordingly obtained. Therefore, when a problem statement, a first preset instruction, and a domain knowledge corresponding to the problem statement are inputted into the intermediate data analysis model, it can accurately output the SQL statement and data statistical analysis function corresponding to the problem statement.

[0085] In certain circumstances, the loss function may not converge even after extensive iterative training. To prevent the iterative training of the initial data analysis model from continuing indefinitely, when it is determined that the loss function of the initial data analysis model obtained after the current iteration of training has not converged, any one of the following three methods may be adopted.

[0086] In a first method, if the loss function of the initial data analysis model has not converged, block 92 includes determining whether a cumulative training duration of the initial data analysis model exceeds a first predetermined cumulative training duration threshold. If the cumulative training duration has exceeded the first predetermined cumulative training duration threshold, it indicates that the iterative training duration has met the requirements. At this point, the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the fine-tuned or intermediate data analysis model. If the cumulative training duration has not exceeded the first predetermined cumulative training duration threshold, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training are conducted based on the optimized and adjusted initial data analysis model.

[0087] A second method is based on a modification of the first method. Particularly, if the loss function of the initial data analysis model has not converged, block 92 includes determining whether a cumulative training iteration count of the initial data analysis model exceeds a first predetermined cumulative training iteration count threshold. If the cumulative training iteration count has exceeded the first predetermined cumulative training iteration count, it indicates that

the iterative training count has met the requirements. At this point, the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the fine-tuned or intermediate data analysis model. If the cumulative training iteration count has not exceeded the first predetermined cumulative training iteration count, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training is conducted based on the optimized and adjusted initial data analysis model.

[0088] In a third method, if the loss function of the initial data analysis model has not converged, block 92 includes validating the data analysis model based on first validation datasets to obtain the corresponding accuracy, precision, and recall rates of the data analysis model. If the accuracy, precision, and recall rates of the data analysis model all meet first predetermined validation requirements, then the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the fine-tuned or intermediate data analysis model. If any one of the accuracy, precision, and recall rates of the data analysis model does not meet the first predetermined validation requirements, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training are conducted based on the optimized and adjusted initial data analysis model.

[0089] Each first validation dataset includes first validation model inputs and first validation model outputs. The first validation model inputs include a validation problem statement, a first validation instruction, and a validation domain knowledge corresponding to the validation problem statement. The first validation model outputs include a validation SQL statement corresponding to the validation problem statement and a validation data statistical analysis function.

[0090] Block 92 may be performed by using the computer processor at the data analysis device, e.g. Figure 4, upon instruction by the data processing application, to perform training of an initial data analysis model to obtain an intermediate data analysis model, including the above-mentioned steps.

[0091] Block 93 includes, using the second training datasets, performing iterations of training of the intermediate data analysis model to obtain a trained data analysis model. In other words, the iterations of training include using the second training model inputs as inputs of the intermediate data analysis model and using the second training model outputs corresponding to the second training model inputs as outputs of the intermediate data analysis model to iteratively train the intermediate data analysis model.

[0092] Block 93 includes, after each iteration of training, determining whether a loss function of the intermediate data analysis model converges. If the loss function does not converge, the model parameters of the initial data analysis model are optimized and adjusted according to

the loss function, and the next iteration of training is performed based on the optimized and adjusted data analysis model. If the loss function converges, e.g., after multiple iterations of training, the data analysis model obtained after completion of this latest iteration training is determined as the trained data analysis model.

[0093] It should be noted that since multiple second training model inputs, e.g., training data statistical analysis results and a second training preset instruction, are used as inputs of the data analysis model, and the second training model outputs, e.g., training abnormal analysis result and training trend analysis results corresponding to the training data statistical analysis results, are used as outputs of the data analysis model to iteratively train the data analysis model, trained data analysis model is accordingly obtained. Therefore, when a data statistical analysis result and a second preset instruction are inputted into the trained data analysis model, it can accurately output the anomaly analysis result and trend analysis result corresponding to the data statistical analysis result.

[0094] In certain circumstances, the loss function may not converge even after extensive iterative training. To prevent the iterative training of the intermediate data analysis model from continuing indefinitely, when it is determined that the loss function of the intermediate data analysis model obtained after the current iteration of training has not converged, any one of the following three methods may be adopted.

[0095] In a first method, if the loss function of the intermediate data analysis model has not converged, block 93 includes determining whether a cumulative training duration of the intermediate data analysis model exceeds a second predetermined cumulative training duration threshold. If the cumulative training duration has exceeded the second predetermined cumulative training duration threshold, it indicates that the iterative training duration has met the requirements. At this point, the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the trained data analysis model. If the cumulative training duration has not exceeded the second predetermined cumulative training duration threshold, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training are conducted based on the optimized and adjusted intermediate data analysis model.

[0096] A second method is based on a modification of the first method. Particularly, if the loss function of the intermediate data analysis model has not converged, block 93 includes determining whether a cumulative training iteration count of the intermediate data analysis model exceeds a second predetermined cumulative training iteration count threshold. If the cumulative training iteration count has exceeded the second predetermined cumulative training iteration count, it indicates that the iterative training count has met the requirements. At this point, the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the trained data analysis model. If the cumulative

training iteration count has not exceeded the second predetermined cumulative training iteration count, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training is conducted based on the optimized and adjusted intermediate data analysis model.

[0097] In a third method, if the loss function of the intermediate data analysis model has not converged, block 93 includes validating the data analysis model based on second validation datasets to obtain the corresponding accuracy, precision, and recall rates of the data analysis model. If the accuracy, precision, and recall rates of the data analysis model all meet first predetermined validation requirements, then the iterative training can be stopped, and the data analysis model obtained after this iteration of training can be determined as the trained data analysis model. If any one of the accuracy, precision, and recall rates of the data analysis model does not meet the second predetermined validation requirements, the model parameters are optimized and adjusted based on the loss function, and subsequent iteration(s) of training are conducted based on the optimized and adjusted intermediate data analysis model.

[0098] Each second validation dataset includes second validation model inputs and second validation model outputs. The second validation model inputs include a validation data statistical analysis result and a second validation preset instruction. The second validation model outputs include a validation anomaly analysis result and a validation trend analysis result corresponding to the validation data statistical analysis result.

[0099] It is to be noted that the trained data analysis model eventually obtained through iterative training of the initial data analysis model based on the first training datasets and the second training datasets has two functions. Firstly, when a problem statement is inputted into the trained data analysis model, it can accurately output the SQL statement and data statistical analysis function corresponding to the problem statement. Compared to a data processing application that generates an SQL statement corresponding to the problem statement based on the problem and selects the data statistical analysis function corresponding to the problem from pre-stored multiple data statistical analysis functions, the trained data analysis model can quickly determine the SQL statement and data statistical analysis function corresponding to the problem statement. Secondly, when a data statistical analysis result is inputted into the trained data analysis model, it can accurately output the anomaly analysis result and trend analysis result corresponding to the data statistical analysis result, thereby enabling quick and accurate data processing and analysis of the retrieved data based on the trained data analysis model.

[0100] Due to the complex structure and ultra-large scale of parameters of the large model, its size is very large. Therefore, iteratively training a first large model based on first training datasets results in a Model A with the aforementioned first function, while iteratively training a

second large model based on second training datasets results in a Model B with the aforementioned second function. The combined size of Model A and Model B is significantly larger than the above-mentioned single data analysis model, and the amount of video memory resources occupied by Model A and Model B in electronic devices is also significantly greater than that of the above-mentioned single data analysis model. Additionally, although small models have a simple structure and contain fewer parameters, the accuracy of the output results of trained small models is much lower than that of trained large models. Therefore, iteratively training a first small model based on the first training datasets results in a Model C with the above-mentioned first function, while iteratively training a second small model based on the second training datasets results in a Model D with the aforementioned second function. The accuracy of the output results of Model C and Model D is much lower than that of the above-mentioned single data analysis model.

[0101] Block 93 may be performed by using the computer processor at the data analysis device, e.g. Figure 4, upon instruction by the data processing application, to perform training of an intermediate data analysis model to obtain a trained data analysis model, including the above-mentioned steps.

[0102] To improve the accuracy of the output results of the trained data analysis model, when a number or cumulative count of the iterations of data processing and analysis exceeds a predetermined cumulative count of iterations of data processing and analysis, retraining of the trained data analysis model may be performed using first precedent datasets and second precedent datasets. Particularly, the flowchart of Figure 3 may include block 600 which may include iteratively retraining the trained data analysis model to obtain a retrained data analysis model, and updating or replacing the trained data analysis model with the retrained data analysis model. Block 600 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to retrain the trained data analysis model to obtain a retrained data analysis model and updating or replacing the trained data analysis model with the retrained data analysis model.

[0103] Particularly, block 600 includes using first precedent datasets, performing iterations of retraining of a trained data analysis model to obtain an intermediately-retrained data analysis model (block 602); using second precedent datasets, performing iterations of retraining of the intermediately-retrained data analysis model to obtain a retrained data analysis model (block 603); and using the retrained data analysis model, updating the trained data analysis model (block 603).

[0104] Block 600 may further include block 601 which includes obtaining a plurality of first precedent datasets and a plurality of second precedent datasets from prior iterations of block 100 to block 400, and storing them in a database. Particularly, each first precedent dataset includes first precedent model inputs and first precedent model outputs. The first precedent

model inputs include the problem statement, the first preset instruction, and the domain knowledge. The first precedent model outputs include the data query, and the data statistical analysis function. Each second precedent dataset includes second precedent model inputs and second precedent model outputs. The second precedent model inputs include the data statistical analysis result, and the second preset instruction. The second precedent model outputs include the anomaly analysis and the trend analysis.

[0105] Block 601 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to obtain the first precedent datasets and the second precedent datasets, and storing them in a database.

[0106] Block 602 includes, using the first precedent datasets, performing iterations of retraining of a trained data analysis model to obtain an intermediately-retrained data analysis model. In other words, the iterations of training include using the first precedent model inputs as inputs of the trained data analysis model and using the first precedent model outputs corresponding to the first precedent model inputs as outputs of the trained data analysis model to iteratively retrain the trained data analysis model.

[0107] Block 602 includes, after each iteration of retraining, determining whether a loss function of the trained data analysis model converges. If the loss function does not converge, the model parameters of the trained data analysis model are optimized and adjusted according to the loss function, and the next iteration of training is performed based on the optimized and adjusted trained data analysis model. If the loss function converges, e.g., after multiple iterations of retraining, the data analysis model obtained after completion of this latest iteration retraining is determined as the intermediately-retrained data analysis model.

[0108] In certain circumstances, the loss function may not converge even after extensive iterative retraining. To prevent the iterative retraining of the trained data analysis model from continuing indefinitely, when it is determined that the loss function of the trained data analysis model obtained after the current iteration of retraining has not converged, any one of the three methods, as described in relation to block 92, may be adopted.

[0109] Block 602 may be performed by using the computer processor at the data analysis device, e.g. Figure 4, upon instruction by the data processing application, to perform retraining of a trained data analysis model, using first precedent datasets, to obtain an intermediately-retrained data analysis model, including the above-mentioned steps.

[0110] Block 603 includes, using the second precedent datasets, performing iterations of retraining of the intermediately-retrained data analysis model to obtain a retrained data analysis model. In other words, the iterations of retraining include using the second precedent model inputs as inputs of the intermediately-retrained data analysis model and using the second precedent model outputs corresponding to the second precedent model inputs as

outputs of the intermediately-retrained data analysis model to iteratively retrain the intermediately-retrained data analysis model.

[0111] Block 603 includes, after each iteration of retraining, determining whether a loss function of the intermediately-retrained data analysis model converges. If the loss function does not converge, the model parameters of the intermediately-retrained data analysis model are optimized and adjusted according to the loss function, and the next iteration of training is performed based on the optimized and adjusted data analysis model. If the loss function converges, e.g., after multiple iterations of training, the data analysis model obtained after completion of this latest iteration training is determined as the retrained data analysis model.

[0112] In certain circumstances, the loss function may not converge even after extensive iterative retraining. To prevent the iterative retraining of the intermediately-retrained data analysis model from continuing indefinitely, when it is determined that the loss function of the intermediately-retrained data analysis model obtained after the current iteration of retraining has not converged, any one of the three methods, as described in relation to block 93, may be adopted.

[0113] Block 603 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to perform retraining of the intermediately-retrained data analysis model, using second precedent datasets, to obtain a retrained data analysis model, including the above-mentioned steps.

[0114] Block 604 includes, using the retrained data analysis model, updating the trained data analysis model. Accordingly, the trained data analysis model is replaced by the retrained data analysis model which is used for subsequent iterations of data processing and analysis.

[0115] Block 604 may be performed by using the computer processor at the data analysis device, e.g., Figure 4, upon instruction by the data processing application, to update or replace the trained data analysis model with the retrained data analysis model.

[0116] Figure 4 shows a schematic diagram of a data processing and analysis system which comprises at least one terminal device 10, at least one data analysis device 20, and at least one database 30. Each terminal device 10 may be a user electronic or computing device which comprises at least one processor 11 which is communicably coupled to at least one memory 12, at least one display device 13, and an input device 14. The display device 13 may be integrated with or separate from the terminal device 10. The input device 14 may receive requests and/or instructions from a user of the terminal device 10. The data analysis device 20 may be an electronic or computing device, e.g., computer, server, which comprises at least one processor 21 which is communicably coupled to at least one memory 22. The data analysis device 20 may include a data analysis application as computer-executable instructions stored in the memory 22 and executed by the processor 21. The data analysis device 20 is communicably coupled to the terminal device 10 and the database 30.

[0117] The above-mentioned processor may be any type of computer processor, such as a Central Processing Unit (CPU), a Graphics Processing Unit (GPU), a quantum processor, a machine learning processor, an artificial intelligence processor, a neural network processor, an artificial intelligence accelerator, an application specific integrated circuit (ASIC), an analog signal processor, an analog computer, a microprocessor, a digital signal processor, a field programmable gate array (FPGA), a tensor processing unit (TPU), a data processing unit (DPU), etc., or other device configured to execute code or computer-executable instructions to perform or implement the flowcharts, algorithms, processes, or operations detailed herein. The memory may be a non-transitory computer readable medium or machine readable medium for storing or comprising the code or computer-executable instructions. Examples include, but are not limited to, volatile memory, e.g., DRAM), nonvolatile memory, e.g., read-only memory (ROM), high bandwidth memory (HBM), flash memory, solid state memory, and/or a hard drive).

[0118] It is to be appreciated that the flow charts showing logic flows are representative of exemplary methodologies for performing novel aspects of the invention. While, for purposes of simplicity of explanation, the one or more methodologies shown herein are shown and described as a series of acts, those skilled in the art will understand and appreciate that the methodologies are not limited by the order of acts. Some acts may, in accordance therewith, occur in a different order and/or concurrently with other acts from that shown and described herein. For example, those skilled in the art will understand and appreciate that a methodology could alternatively be represented as a series of interrelated states or events, such as in a state diagram. Moreover, not all acts illustrated in a methodology may be required for a novel implementation.

[0119] The flow charts showing logic flow may be implemented in software, firmware, hardware, or any combination thereof. In software and firmware embodiments, the logic flow may be implemented by computer executable instructions or code stored on a non-transitory computer-readable medium or machine readable medium, such as an optical, magnetic or semiconductor storage. The computer-executable instructions are configured to direct at least one computer processor to perform the logic flow. The embodiments are not limited in this context.

[0120] It is to be understood that the embodiments and features described above should be considered exemplary and not restrictive. Many other embodiments will be apparent to those skilled in the art from consideration of the specification and practice of the invention. Furthermore, certain terminology has been used for the purposes of descriptive clarity, and not to limit the disclosed embodiments of the invention.

CLAIMS

[Claim 1] A computer-implemented data processing and analysis method, comprising:
performing iterations of data processing and analysis, wherein each iteration comprises:
receiving a problem statement sent by a terminal device, wherein the problem statement defines a data retrieval requirement corresponding thereto;
generating a data query based on the data retrieval requirement, and determining a data statistical analysis function based on the problem statement;
retrieving data based on the data query;
generating a data statistical analysis result based on the retrieved data and the data statistical analysis function;
generating, by a trained data analysis model, an anomaly analysis and a trend analysis based on the data statistical analysis result and a second instruction; and
providing the data statistical analysis result, the anomaly analysis, and the trend analysis to the terminal device or an other device.

[Claim 2] The method of claim 1, wherein generating the data query based on the data retrieval requirement and determining the data statistical analysis function based on the problem statement includes:

generating, using the trained data analysis model, the data query based on the data retrieval requirement and a first instruction; and

generating, using the trained data analysis model, the data statistical analysis function based on the problem statement and the first instruction.

[Claim 3] The method of claim 1 or claim 2, further comprising:

using first training datasets, performing iterations of training of an initial data analysis model to obtain an intermediate data analysis model, wherein each first training dataset includes first training model inputs and first training model outputs, wherein the first training model inputs include a training problem statement, and a first training instruction, wherein first training model outputs include a training data query, and a training data statistical analysis function; and

using second training datasets, performing iterations of training of the intermediate data analysis model to produce the trained data analysis model, wherein each second training dataset includes second training model inputs and second training model outputs,

wherein the second training model inputs include a training data statistical analysis result, and a second training instruction, wherein the second training model outputs include a training anomaly analysis and a training trend analysis.

[Claim 4] The method of claim 3, wherein performing the iterations of training of the initial data analysis model to obtain the intermediate data analysis model includes:

after performing each iteration of training:

determining a loss function of the initial data analysis model;

determining the loss function of the initial data analysis model is non-convergent and adjusting model parameters of the initial data analysis model;

wherein the method further comprises:

stop performing the iterations of training of the initial data analysis model if a cumulative training duration of the initial data analysis model exceeds a first predetermined cumulative training duration threshold, or a cumulative training iteration count of the initial data analysis model exceeds a first predetermined cumulative training iteration count threshold, or a validation of the initial data analysis model exceeds a first predetermined validation requirement; and

wherein performing the iterations of training of the intermediate data analysis model to produce the trained data analysis model includes:

after performing each iteration of training:

determining a loss function of the intermediate data analysis model;

determining the loss function of the intermediate data analysis model is non-convergent and adjusting model parameters of the intermediate data analysis model;

wherein the method further comprises:

stop performing the iterations of training of the intermediate data analysis model if a cumulative training duration of the intermediate data analysis model exceeds a second predetermined cumulative training duration threshold, or a cumulative training iteration count of the intermediate data analysis model exceeds a second predetermined cumulative training iteration count threshold, or a validation of the intermediate data analysis model exceeds a second predetermined validation requirement.

[Claim 5] The method of claim 2, further comprising:

retraining the trained data analysis model, when a cumulative count of the iterations of data processing and analysis exceeds a predetermined cumulative count of iterations of data processing and analysis, to obtain a retrained data analysis model including:

using first precedent datasets obtained from performing the iterations of data processing and analysis, performing iterations of retraining of the trained data analysis model to obtain an intermediately-retrained data analysis model, wherein each first precedent dataset includes first precedent model inputs and first precedent model outputs, wherein the first precedent model inputs include the problem statement, and the first instruction, wherein first precedent model outputs include the data query, and the data statistical analysis function;

using second precedent datasets obtained from performing iterations of data processing and analysis, performing iterations of retraining on the trained data analysis model to obtain an intermediately-retrained data analysis model wherein each second precedent dataset includes second precedent model inputs and second precedent model outputs, wherein the second precedent model inputs include the data statistical analysis result, and the second instruction, wherein the second precedent model outputs include the anomaly analysis and the trend analysis; and

using the retrained data analysis model, updating the trained data analysis model.

[Claim 6] The method of claim 5, further comprising:

receiving a correction request having at least one of the group consisting of a corrected data query, a corrected data statistical analysis function, a correct data statistical analysis result, a corrected anomaly analysis, and a corrected trend analysis;

storing the correction request; and

retraining the trained data analysis model based on the correction request, wherein at least one of the first precedent datasets or at least one of the second precedent datasets is based on the correction request.

[Claim 7] The method of claim 1, wherein generating the data query based on the data retrieval requirement includes:

generating a Structured Query Language (SQL) statement; and

retrieving the data from the database based on the generated SQL statements.

[Claim 8] The method of claim 1, wherein generating the data statistical analysis result based on the retrieved data and the data statistical analysis function includes:

storing a plurality of sample problem statements and a plurality of sample data statistical analysis functions corresponding to the sample problem statements;

successfully matching the problem statement to one of the sample problem statements;

determining a corresponding sample data statistical analysis function of the matched sample problem statement as the data statistical analysis function corresponding to the problem statement.

[Claim 9] The method of claim 1, wherein generating the data query based on the data retrieval requirement includes:

determining a domain knowledge of the problem statement; and

generating, using the data analysis model, the data query based on the problem statement based on the domain knowledge and the problem statement.

[Claim 10] The method of claim 9, wherein determining the domain knowledge of the problem statement includes:

determining similarity scores of sample problem statements relative to the problem statement, determining a corresponding one of the sample problem statements as having a similarity score higher than a predetermined score threshold, determining a domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement.

[Claim 11] The method of claim 9, wherein determining the domain knowledge of the problem statement includes:

determining keywords of the problem statement, determining matching counts of the sample problem statements relative to the problem statement, determining a corresponding one of the sample problem statements as having a highest matching count, determining a

domain knowledge of the corresponding one of the sample problem statements as the domain knowledge of the problem statement.

[Claim 12] The method of claim 1, wherein generating the data statistical analysis result based on the retrieved data and the data statistical analysis function includes:

- substituting the retrieved data into the data statistical analysis function; and
- performing the data statistical analysis function to generate the data statistical analysis result.

[Claim 13] A non-transitory computer-readable medium comprising computer-executable instructions configured to direct at least one computer processor to perform the method according to any one of claim 1 to claim 12.

[Claim 14] A data processing and analysis device comprising:

- at least one memory for storing computer-executable instructions; and
- at least one processor communicably coupled to the at least one memory, and configured to:
 - execute the computer-executable instructions to perform the method according to any one of claim 1 to claim 12.